Hindsight judgement on ambiguous episodes of suspected infection in critically ill children: poor consensus amongst experts?

Fischer, Joachim E ; Seifarth, Federico G ; Baenziger, Oskar ; Fanconi, Sergio ; Nadal, David

Abstract: Few episodes of suspected infection observed in paediatric intensive care are classifiable without ambiguity by a priori defined criteria. Most require additional expert judgement. Recently, we observed a high variability in antibiotic prescription rates, not explained by the patients' clinical data or underlying diseases. We hypothesised that the disagreement of experts in adjudication of episodes of suspected infection could be one of the potential causes for this variability. During a 5-month period, we included all patients of a 19-bed multidisciplinary, tertiary, neonatal and paediatric intensive care unit, in whom infection was clinically suspected and antibiotics were prescribed (n=183). Three experts (two senior ICU physicians and a specialist in infectious diseases) were provided with all patient data, laboratory and microbiological findings. All experts classified episodes according to a priori defined criteria into: proven sepsis, probable sepsis (negative cultures), localised infection and no infection. Episodes of proven viral infection and incomplete data sets were excluded. Of the remaining 167 episodes, 48 were classifiable by a priori criteria (n=28 proven sepsis, n= 20 no infection). The three experts only achieved limited agreement beyond chance in the remaining 119 episodes (kappa = 0.32, and kappa = 0.19 amongst the ICU physicians). The kappa is a measure of the degree of agreement beyond what would be expected by chance alone, with 0 indicating the chance result and 1 indicating perfect agreement. Conclusion: agreement of specialists in hindsight adjudication of episodes of suspected infection is of questionable reliability

DOI: https://doi.org/10.1007/s00431-002-0959-z

Eur J Pediatr (2003) 162: 840–843
DOI 10.1007/s00431-002-0959-z

ORIGINAL PAPER

**Joachim E. Fischer · Federico G. Seifarth
Oskar Baenziger · Sergio Fanconi · David Nadal**

Hindsight judgement on ambiguous episodes of suspected infection in critically ill children: poor consensus amongst experts?

Received: 14 August 2001 / Accepted: 25 February 2002 / Published online: 2 October 2003

Abstract Few episodes of suspected infection observed in paediatric intensive care are classifiable without ambiguity by a priori defined criteria. Most require additional expert judgement. Recently, we observed a high variability in antibiotic prescription rates, not explained by the patients' clinical data or underlying diseases. We hypothesised that the disagreement of experts in adjudication of episodes of suspected infection could be one of the potential causes for this variability. During a 5-month period, we included all patients of a 19-bed multidisciplinary, tertiary, neonatal and paediatric intensive care unit, in whom infection was clinically suspected and antibiotics were prescribed ($n = 183$). Three experts (two senior ICU physicians and a specialist in infectious diseases) were provided with all patient data, laboratory and microbiological findings. All experts classified episodes according to a priori defined criteria into: proven sepsis, probable sepsis (negative cultures), localised infection and no infection. Episodes of proven viral infection and incomplete data sets were excluded. Of the remaining 167 episodes, 48 were classifiable by a priori criteria ($n = 28$ proven sepsis, $n = 20$ no infection). The three experts only achieved limited agreement beyond chance in the remaining 119 episodes (kappa = 0.32, and kappa = 0.19 amongst the ICU physicians). The kappa is a measure of the degree of agreement beyond what would be expected by chance alone, with 0 indicating the chance result and 1 indicating perfect agreement. *Conclusion*: Agreement of specialists in hindsight adjudication of episodes of suspected infection is of questionable reliability.

Keywords Infection · Interobserver-agreement
Outcome adjudication · Paediatric intensive care

J. E. Fischer (✉) · F. G. Seifarth · O. Baenziger · S. Fanconi
Department of Neonatology and Paediatric Intensive Care,
University Children's Hospital,
Steinwiesstrasse 75, 8032 Zurich,
Switzerland
E-mail: joachim.fischer@kispi.unizh.ch
Tel.: +41-1-2667751
Fax: +41-1-2667171

S. Fanconi
Department of Paediatrics,
CHUV, Lausanne, Switzerland

D. Nadal
Division of Infectious Diseases,
University Children's Hospital,
Zurich, Switzerland

Introduction

A daily task in paediatric intensive care units is to assess patients as to the presence or absence of infection. For every patient, a decision must be made: to do nothing, to order tests, to order a formal sepsis work-up, to start, stop or change antibiotic treatment. The physician responsible for this decision must integrate a large number of variables, including the history, the laboratory findings, the clinical presentation and the observations made by nurses or others, including parents [9]. When a decision has to be made, a high degree of uncertainty often remains. Because of the high risks associated with untreated infection, physicians have a low threshold to prescribe "rule-out" antibiotics. Usually, antibiotics are discontinued after 48 h if cultures remain negative and alternative explanations for the clinical findings are likely. Thus, the initial decision to give antibiotics is repetitively scrutinised in order to minimise inappropriate prescription. Recently, we observed a high variability in prescription rates in a paediatric intensive care unit, which was not accounted for by patient data [4]. We speculated that physicians' uncertainty in judgement on ambiguous cases contributes to this variability.

In ambiguous episodes of suspected infection, senior clinicians base their adjudication on clinical signs, history, laboratory parameters, the time course of events, and their own experience. An initial treatment decision

may be scrutinised during subsequent ward rounds. These reviews of initial decisions are based on reviewing the evolving clinical and laboratory data. In studies on diagnostic markers of infection, hindsight review of purpose-designed charts is the standard method to adjudicate outcome [1, 6,7]. To investigate the reliability and validity of hindsight judgement in ambiguous cases of suspected infection, we provided three senior clinicians, who were blinded to the judgement of each other, with all available data from an inception cohort of cases of suspected infection in critically ill children. We asked the clinicians to decide on the most likely of five possible diagnoses, which each require different treatment: sepsis, probable sepsis with negative blood cultures, localised infection, viral infection or absent infection.

Patients and methods

Patients

The study comprised consecutive episodes of suspected infection occurring in patients admitted to a level-3, multidisciplinary, neonatal and paediatric intensive care unit (PICU). The 19-bed PICU is the tertiary referral centre for Eastern and Southern Switzerland. The unit provides treatment for children with severe medical conditions or trauma, post-operative care after cardiac surgery or any major paediatric or neonatal surgery, and cares for outborn neonates with critical illness.

Study design

During a 5-month period, we included all patients with an episode of suspected infection. Clinical suspicion of infection was defined as (1) change in the prescription of antibiotics (new antibiotics or change from prophylaxis to treatment), (2) an explicit statement in the patient records that infection was suspected, (3) the initiation of a diagnostic work-up (blood cultures, local cultures, white blood cell count, differential count and determination of C-reactive protein). Thus, we only included patients who received antibiotic treatment for suspected infection. A research assistant (F.G.S) collected data relevant to the adjudication of infection, which are not systematically collected in the patients' charts, into a purpose-designed database.

Three senior physicians (two consultants, J.E.F. and O.B. and one full-time specialist in infectious diseases, D.N.) were provided with printouts from the database, copies of the patients' charts, the discharge letters, and all microbiological and laboratory data. The three physicians were asked to adjudicate each episode of suspected infection into one of the following categories according to criteria published elsewhere [4]: culture-proven sepsis, probable sepsis with negative blood cultures, localised infection, viral infection or infection unlikely. The latter episodes were defined as absent infection provided antibiotics were successfully discontinued within 48 h. All experts agreed on the adjudication criteria. They were blinded to the judgment of each other.

In order to simulate the case of an episode of suspected infection that can be unanimously classified based on simple criteria, we provided a fifth year medical student, who did not have any clinical experience, with the same criteria as the experts. The medical student was asked to identify all episodes of culture-proven sepsis and episodes of absent infection. After excluding episodes of antigen-proven viral illness, we determined which episodes were classified as culture-proven sepsis or as absent infection by all experts and by the student. These episodes were regarded as episodes being classifiable on hindsight. The remaining episodes were regarded as ambiguous. For these remaining ambiguous episodes we determined the agreement beyond chance amongst experts (Fig. 1). Experts were informed that the purpose of the judgment was to achieve unbiased outcome adjudication in a concurrent study on new markers of infection. The intent of the present analyses was not disclosed until adjudication was completed. Expert adjudication was performed within 2 months after the last patient had been enrolled.

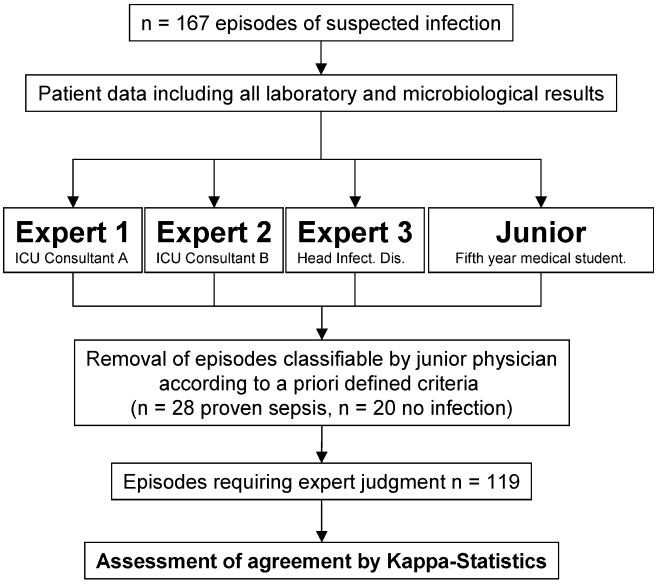

Fig. 1 Study flow chart

Definitions

A priori defined criteria for culture-proven sepsis comprised: clinical suspicion, two or more of the clinical signs outlined in [4], a C-reactive protein level >20 mg/l, a left shift with a ratio of immature to total neutrophils exceeding 0.5 and a positive blood culture for a pathogen other than coagulase-negative Staphylococci or two positive blood cultures obtained from separate sites growing coagulase-negative Staphylococci. Criteria for no infection comprised: clinical suspicion, negative cultures from local sites and from blood, all determinations of C-reactive protein yield results <20 mg/l, ratio of immature to total neutrophils <0.4 and antibiotics discontinued within 48 h [4].

Data analysis

For each patient only the first episode was entered into the calculation. Agreement beyond chance was calculated as Cohen's kappa [5], which is a measure of the agreement beyond chance. A kappa of 0.8–1.0 is considered as almost perfect agreement, a kappa of 0.5 indicates that the adjudicators achieved 50% of the possible agreement beyond chance. Kappa values ranging from 0 to 0.2 suggest that the agreement among raters is not better than chance [8]. Calculations were performed using the STATA software program (STATA Corp., College Station, Texas, USA).

Results

During the study period, 328 patients were admitted. The median age was 0.67 years (mean = 1.90 years, standard deviation = 2.44 years, 10th percentile 0 years,

90th percentile 9.6 years). Bacterial infection was suspected in 183 patients and all were started on antibiotics. Of these, 70 (38%) had undergone major surgery or were trauma patients. C-reactive protein levels exceeded the cutoff of 20 mg/l at diagnostic work-up in 116 (63%) patients. Nine patients had viral infections, and seven datasets were incomplete, leaving 167 patients for the analysis (Table 1). Of these, 39 (23%) patients showed positive blood cultures. In 11 patients, one positive culture grew coagulase-negative Staphylococci, the remaining 28 positive cultures grew *S. epidermidis* from two independently obtained bottles or other pathogens. These 28 (17%) patients satisfied the a priori defined criteria for culture-proven sepsis. A further 20 (11%) patients met the criteria for no infection. The remaining 119 (71%) episodes were regarded as ambiguous. Table 1presents patient characteristics and the results from episode adjudication. The frequency of outcome categories adjudicated by each expert was similar.

If all 167 episodes were considered, the combined agreement amongst experts from database and chart review beyond chance was moderate ($\kappa = 0.54$), with almost perfect agreement as to episodes of proven sepsis ($\kappa = 0.92$), and slight agreement beyond chance regarding episodes of probable sepsis ($\kappa = 0.18$). However, when those 48 episodes classifiable on hindsight according to the a priori defined criteria were removed, expert agreement on the remaining 119 episodes (71%) was only fair ($\kappa = 0.32$). The two ICU physicians agreed on the adjudication of 41% of these episodes ($\kappa = 0.19$), and the agreement of the two ICU specialists with the expert on infectious diseases was 50% and 62%, ($\kappa = 0.31$ and $\kappa = 0.46$ respectively). Similar results were obtained from calculating the agreement for ''any bacterial infection'' versus ''no infection'' (data not shown).

Table 1 Patient characteristics and outcome adjudication. Clinicians A and B: full-time intensive care consultants; clinician C: consultant for infectious diseases

	All episodes	Episodes requiring expert judgment
Main reason for admission		
Suspected infection (all age groups)	63 (37.8%)	39 (32.8%)
Cardiac surgery	35 (21.0%)	29 (24.4%)
Other surgery or trauma	24 (14.4%)	16 (13.4%)
Medical conditions	16 (9.6%)	13 (10.9%)
Neonatal conditions	29 (17.4%)	22 (18.5%)
Outcome adjudication		
Proven sepsis		
Clinician A	41 (24.6%)	13 (10.9%)
Clinician B	47 (29.0%)	19 (16.0%)
Clinician C	45 (26.9%)	17 (14.3%)
Probable sepsis		
Clinician A	23 (13.8%)	23 (19.3%)
Clinician B	32 (19.8%)	32 (26.9%)*
Clinician C	14 (8.4%)	14 (11.8%)*
Localized infection		
Clinician A	45 (26.9%)	45 (37.8%)
Clinician B	39 (23.3%)	39 (32.7%)
Clinician C	42 (25.1%)	42 (35.3%)
Infection unlikely		
Clinician A	58 (34.7%)	38 (31.9%)
Clinician B	49 (29.3%)	29 (24.4%)**
Clinician C	66 (39.5%)	46 (38.7%)**

*$P = 0.0049$, two-sided Fisher's exact test, these results are not significant if a Bonferroni correction is applied for multiple comparisons (12 comparisons, required $P = 0.00417$)
**$P = 0.025$, two-sided Fisher's exact test, these results are not significant if a Bonferroni correction is applied for multiple comparisons (12 comparisons, required $P = 0.00417$)

Discussion

In this study we examined the inter-rates reliability of a usual method to review treatment decisions in critically ill newborns and children, the hindsight case review. In critically ill patients developing symptoms compatible with infection, clinicians often favour prescription of ''rule-out'' antibiotics in lieu of awaiting the results from microbiological cultures. In most units, decisions for ''rule-out'' therapy are subsequently scrutinised. Experts meet at bedside during the subsequent ward rounds, the responsible physician presents the case, and the previous treatment decision is reviewed. As the illness episode evolves and more data become available, clinicians achieve agreement as to the most likely cause of the patients' symptoms. Based on our recent observation on a high variability in prescription rates, which was not explained by patient characteristics [4], we questioned the validity of this hindsight procedure, which is also used for outcome adjudication in clinical studies [1, 6,7].

The main finding from this study is that approximately 33% of all episodes of suspected infection, in particular those with positive blood cultures, can be unambiguously classified according to pre-specified criteria – even by junior physicians. However, the remaining 66% of all episodes remain ambiguous and even seasoned experts only achieve fair agreement in adjudication. Disturbingly, the agreement of the two clinicians sharing the responsibility for the same patients in the same unit was little better than chance ($\kappa = 0.19$) [8].

To what extent may these findings be generalised and what are the clinical implications? Antibiotic prescription in ambiguous episodes is a potential source of the observed variability: if one clinician determines that an episode is not due to bacterial infection, while the other assumes probable sepsis with negative cultures – highly divergent prescription rates will result. By using a design where each clinician was blinded to the decision of the other, we showed that the individual adjudication is often arbitrary, although all three clinicians arrived at similar overall rates of potential aetiologies. The proportion and the case mix of ultimately ambiguous cases in this study was similar to studies investigating new parameters of infection [7]. Our findings replicate other studies showing limited agreement in clinical judgment on potentially ambiguous outcomes [2, 3,10]. Strictly

speaking, our data demonstrate that the post-hoc extraction of information from the charts, even under conditions of a prospectively and purpose-designed data collection, does not provide for reliable adjudication. The most likely reason is that experts differ in their way to extract and weigh the relevant information. It remains unknown, to what extent reviewing the patients at bedside would have improved the agreement.

The physicians' dilemma is to have to make treatment decisions based on the assumed aetiology whilst facing uncertainty. One may ask, what are experts good for if they disagree? We believe a watchful experienced clinician deciding to withhold antibiotic treatment in order to minimise inappropriate prescription may become alerted by small deviations in the clinical course more rapidly than a junior. She or he will then consider the alternatives, e.g. that the patient has culture-negative sepsis. Until improved diagnostic markers become available, the observed uncertainty will continue to give rise to controversy over the presence or absence of infection.

The limitation of this study is that we may have underestimated potential agreement due to some patients simultaneously having more than one of the conditions (e.g. a newborn after cardiac surgery may simultaneously develop a post-surgical systemic inflammatory reaction and be incubating ventilator-associated pneumonia). Expert review of the case at bedside may improve adjudication over the level reported here. However, it is conceivable that decisions made after verbal reporting of the patient findings over the telephone, a frequent situation during night-duty on-call, may result in potential agreement, which is lower than in our investigation. In the adjudication of cases of suspected infection which are not classifiable by stringent a priori defined criteria, the level of expert agreement after reviewing charts is limited.

Acknowledgement The study was supported by a grant from the Alice Bucher Foundation, Lucerne, Switzerland.

References

1. Cook DJ, Walter SD, Cook RJ, Griffith LE, Guyatt GH, Leasa D, Jaeschke RZ, Brun-Buisson C (1998) Incidence of and risk factors for ventilator-associated pneumonia in critically ill patients. Ann Intern Med 129: 433–440
2. Davies HD, Wang EE, Manson D, Babyn P, Shuckett B (1996) Reliability of the chest radiograph in the diagnosis of lower respiratory infections in young children. Pediatr Infect Dis J 15: 600–604
3. Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario (1980) Clinical disagreement: II. How to avoid it and how to learn from one's mistakes. Can Med Assoc J 123: 613–617
4. Fischer JE, Ramser M, Fanconi S (2000) Use of antibiotics in pediatric intensive care and potential savings. Intensive Care Med 26: 959–966
5. Fleiss JL (1981) Statistical methods for rates and proportions. Wiley, New York, pp 217–234
6. Heyland DK, Cook DJ, Marshall J, Heule M, Guslits B, Lang J, Jaeschke R (1999) The clinical utility of invasive diagnostic techniques in the setting of ventilator-associated pneumonia. Canadian Critical Care Trials Group. Chest 115: 1076–1084
7. Kuster H, Weiss M, Willeitner AE, Detlefsen S, Jeremias I, Zbojan J, Geiger R, Lipowsky G, Simbruner G (1998) Interleukin-1 receptor antagonist and interleukin-6 for early diagnosis of neonatal sepsis 2 days before clinical manifestation. Lancet 352: 1271–1277
8. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33: 159–174
9. Sacket DL, Haynes RB, Guyatt GH, Tugwell P (1991) Clinical epidemiology: a basic science for clinical medicine, 2nd edn.. Little, Brown, Boston Toronto London
10. Wipf JE, Lipsky BA, Hirschmann JV, Boyko EJ, Takasugi J, Peugeot RL, Davis CL (1999) Diagnosing pneumonia by physical examination: relevant or relic? Arch Intern Med 159: 1082–1087